(19)日本国特許庁（ＪＰ）　(12)**公開特許公報（Ａ）**　(11)特許出願公開番号
特開2002－373059
（P2002－373059A）
(43)公開日　平成14年12月26日(2002.12.26)

(51)Int.Cl.'		識別記号	FI		テーマコード'(参考)
Ｇ０６Ｆ	3/06	５４０	Ｇ０６Ｆ 3/06	５４０	５Ｂ０１８
	12/16	３２０	12/16	３２０Ｌ	５Ｂ０６５

審査請求　未請求　請求項の数７　ＯＬ　（全 10 頁）

(21)出願番号　特願2001－180202(P2001－180202)

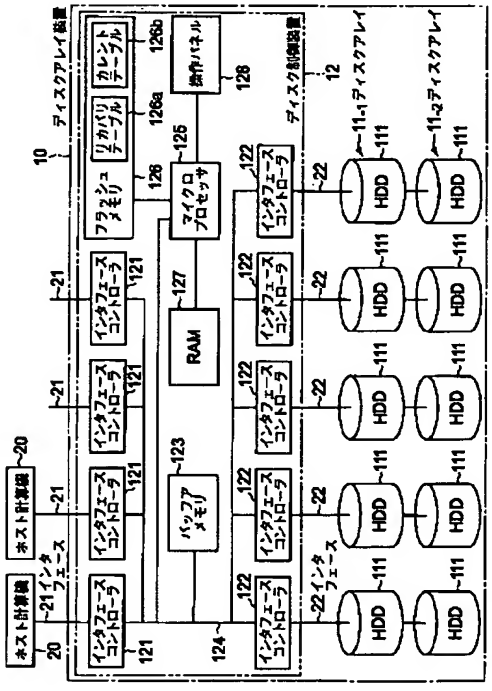
(22)出願日　平成13年６月14日(2001.6.14)

(71)出願人　000003078
株式会社東芝
東京都港区芝浦一丁目１番１号
(72)発明者　笹本　享一
東京都府中市東芝町１番地　株式会社東芝
府中事業所内
(74)代理人　100058479
弁理士　鈴江　武彦　（外６名）

Ｆターム(参考)　5B018 GA06 KA15 MA14
5B065 EA11 EA18

(54)【発明の名称】　ディスクアレイのエラー回復方法、ディスクアレイ制御装置及びディスクアレイ装置

(57)【要約】
【課題】ディスクドライブの多重故障によりディスクアレイが閉塞したとしても、当該ディスクアレイを簡単に使用可能な状態に復旧できるようにする。
【解決手段】ディスクアレイ制御装置１２内のマイクロプロセッサ１２５は、ディスクアレイ１１-i内の複数のＨＤＤ１１１の故障により当該アレイが閉塞する際、閉塞直前のカレントテーブル１２６ｂ上のＲＡＩＤ構成情報をリカバリテーブル１２６ａに保存し、当該アレイの閉塞後に、カレントテーブル１２６ｂ上のＲＡＩＤ構成情報を当該アレイの閉塞後の状態を反映するように更新する。ユーザが操作パネル１２８を操作することでリカバリ要求が与えられた場合、マイクロプロセッサ１２５はリカバリテーブル１２６ａから指定のディスクアレイ１１-iのＲＡＩＤ構成情報を読み出し、当該情報に従ってアレイ１１-iを閉塞直前の状態に戻し、当該情報をカレントテーブル１２６ｂに上書きする。



【特許請求の範囲】

【請求項１】 複数のディスクドライブから構成されるＲＡＩＤ（Redundant Arrays of Inexpensive Disks）構成の少なくとも１つのディスクアレイを備えると共に、現在の前記ディスクアレイの少なくとも構成及び稼働状態を示す第１のＲＡＩＤ構成情報が保存される不揮発性の第１の記憶領域と、前記ディスクアレイが閉塞した際に、その閉塞直前の当該ディスクアレイの少なくとも構成及び稼働状態を示す第２のＲＡＩＤ構成情報が保存される第２の記憶領域とが確保されたディスクアレイ装置におけるディスクアレイのエラー回復方法であって、

前記ディスクアレイを構成する前記複数のディスクドライブのうちの少なくとも２つのディスクドライブが故障した場合に、当該ディスクアレイを閉塞するステップと、

前記ディスクアレイを閉塞するに際し、前記第１の記憶領域に保存されている当該ディスクアレイの閉塞直前の第１のＲＡＩＤ構成情報を前記第２のＲＡＩＤ構成情報として前記第２の記憶領域に保存するステップと、

前記ディスクアレイの閉塞後、当該ディスクアレイの前記第１の記憶領域上の前記第１のＲＡＩＤ構成情報を当該ディスクアレイの閉塞後の状態を反映するように更新するステップと、

前記ディスクアレイ装置に対して閉塞状態にある前記ディスクアレイを使用可能な状態に回復させるリカバリ要求がユーザ操作に従って与えられた場合に、前記第２の記憶領域から、当該リカバリ要求で指定されたディスクアレイの閉塞直前の前記第２のＲＡＩＤ構成情報を読み出すステップと、

前記第２の記憶領域から読み出された前記第２のＲＡＩＤ構成情報に基づき前記ディスクアレイを閉塞直前の状態に強制的に戻すと共に当該第２のＲＡＩＤ構成情報を前記第１のＲＡＩＤ構成情報として前記第１の記憶領域に上書きするステップとを具備することを特徴とするディスクアレイのエラー回復方法。

【請求項２】 前記ディスクアレイを閉塞するに際し、前記第２のＲＡＩＤ構成情報と共に当該ディスクアレイが閉塞した原因を示す情報も前記第２の記憶領域に保存し、

前記リカバリ要求に従って前記ディスクアレイを閉塞直前の状態に強制的に戻す際には、前記第２の記憶領域から、当該ディスクアレイの閉塞直前の前記第２のＲＡＩＤ構成情報に加えて当該ディスクアレイが閉塞した原因を示す情報を読み出すステップと、

読み出された閉塞原因情報がディスクドライブの部分的な障害であるメディアエラーを示している場合、当該メディアエラーの発生していたブロックを代替用のセクタブロックに代替処理するステップとを更に具備し、

前記代替処理後に、前記閉塞直前の前記ディスクアレイの前記第２のＲＡＩＤ構成情報を前記第１のＲＡＩＤ構成情報として前記第１の記憶領域に上書きすることを特徴とする請求項１記載のディスクアレイのエラー回復方法。

【請求項３】 複数のディスクドライブから構成されるＲＡＩＤ（Redundant Arrays of Inexpensive Disks）構成の少なくとも１つのディスクアレイを制御するディスクアレイ制御装置において、

現在の前記ディスクアレイの少なくとも構成及び稼働状態を示す第１のＲＡＩＤ構成情報が保存される第１の記憶領域と、前記ディスクアレイが閉塞した際に、その閉塞直前の当該ディスクアレイの少なくとも構成及び稼働状態を示す第２のＲＡＩＤ構成情報が保存される第２の記憶領域とが確保された不揮発性記憶手段と、

前記ディスクアレイを構成する前記複数のディスクドライブのうちの少なくとも２つのディスクドライブが故障した場合に、当該ディスクアレイを閉塞する手段と、

前記閉塞手段により前記ディスクアレイが閉塞される際に、前記第１の記憶領域に保存されている当該ディスクアレイの閉塞直前の第１のＲＡＩＤ構成情報を前記第２のＲＡＩＤ構成情報として前記第２の記憶領域に保存すると共に、前記ディスクアレイの閉塞後、当該ディスクアレイの前記第１の記憶領域上の前記第１のＲＡＩＤ構成情報を当該ディスクアレイの閉塞後の状態を反映するように更新する閉塞時ＲＡＩＤ構成情報更新手段と、

前記ディスクアレイ装置に対して閉塞状態にある前記ディスクアレイを使用可能な状態に回復させるリカバリ要求がユーザ操作に従って与えられた場合に、前記第２の記憶領域から、当該リカバリ要求で指定されたディスクアレイの閉塞直前の前記第２のＲＡＩＤ構成情報を読み出す手段と、

前記第２の記憶領域から読み出された前記第２のＲＡＩＤ構成情報に基づき前記ディスクアレイを閉塞直前の状態に強制的に戻すと共に当該第２のＲＡＩＤ構成情報を前記第１のＲＡＩＤ構成情報として前記第１の記憶領域に上書きする復旧手段とを具備することを特徴とするディスクアレイ制御装置。

【請求項４】 複数のディスクドライブから構成されるＲＡＩＤ（Redundant Arrays of Inexpensive Disks）構成の少なくとも１つのディスクアレイであって、現在の当該ディスクアレイの少なくとも構成及び稼働状態を示す第１のＲＡＩＤ構成情報が保存される第１の記憶領域と、前記ディスクアレイが閉塞した際に、その閉塞直前の当該ディスクアレイの少なくとも構成及び稼働状態を示す第２のＲＡＩＤ構成情報が保存される第２の記憶領域とが確保された複数のディスクドライブから構成されるディスクアレイを制御するディスクアレイ制御装置であり、

前記ディスクアレイを構成する前記複数のディスクドライブのうちの少なくとも２つのディスクドライブが故障

した場合に、当該ディスクアレイを閉塞する手段と、
前記閉塞手段により前記ディスクアレイが閉塞される際に、前記第１の記憶領域に保存されている当該ディスクアレイの閉塞直前の第１のＲＡＩＤ構成情報を前記第２のＲＡＩＤ構成情報として前記第２の記憶領域に保存すると共に、前記ディスクアレイの閉塞後、当該ディスクアレイの前記第１の記憶領域上の前記第１のＲＡＩＤ構成情報を当該ディスクアレイの閉塞後の状態を反映するように更新する閉塞時ＲＡＩＤ構成情報更新手段と、
前記ディスクアレイ装置に対して閉塞状態にある前記ディスクアレイを使用可能な状態に回復させるリカバリ要求がユーザ操作に従って与えられた場合に、前記第２の記憶領域から、当該リカバリ要求で指定されたディスクアレイの閉塞直前の前記第２のＲＡＩＤ構成情報を読み出す手段と、
前記第２の記憶領域から読み出された前記第２のＲＡＩＤ構成情報に基づき前記ディスクアレイを閉塞直前の状態に強制的に戻すと共に当該第２のＲＡＩＤ構成情報を前記第１のＲＡＩＤ構成情報として前記第１の記憶領域に上書きする復旧手段とを具備することを特徴とするディスクアレイ制御装置。
【請求項５】　前記閉塞時ＲＡＩＤ構成情報更新手段は、前記閉塞手段により前記ディスクアレイが閉塞される際に、前記第２のＲＡＩＤ構成情報と共に当該ディスクアレイが閉塞した原因を示す情報を前記第２の記憶領域に保存するように構成され、
前記リカバリ要求が与えられた場合に前記読み出し手段によって前記第２の記憶領域から読み出された前記閉塞原因情報がディスクドライブの部分的な障害であるメディアエラーを示している場合、当該メディアエラーの発生していたブロックを代替用のセクタブロックに代替処理するセクタブロック代替手段を更に具備することを特徴とする請求項３または請求項４記載のディスクアレイ制御装置。
【請求項６】　複数のディスクドライブから構成されるＲＡＩＤ（Redundant Arrays of Inexpensive Disks）構成の少なくとも１つのディスクアレイと、前記ディスクアレイを制御するディスクアレイ制御装置とを備えたディスクアレイ装置において、
現在の前記ディスクアレイの少なくとも構成及び稼働状態を示す第１のＲＡＩＤ構成情報が保存される第１の記憶領域と、前記ディスクアレイが閉塞した際に、その閉塞直前の当該ディスクアレイの少なくとも構成及び稼働状態を示す第２のＲＡＩＤ構成情報が保存される第２の記憶領域とが確保された不揮発性記憶手段を備えると共に、
前記ディスクアレイ制御装置は、
前記ディスクアレイを構成する前記複数のディスクドライブのうちの少なくとも２つのディスクドライブが故障した場合に、当該ディスクアレイを閉塞する手段と、
前記閉塞手段により前記ディスクアレイが閉塞される際に、前記第１の記憶領域に保存されている当該ディスクアレイの閉塞直前の第１のＲＡＩＤ構成情報を前記第２のＲＡＩＤ構成情報として前記第２の記憶領域に保存すると共に、前記ディスクアレイの閉塞後、当該ディスクアレイの前記第１の記憶領域上の前記第１のＲＡＩＤ構成情報を当該ディスクアレイの閉塞後の状態を反映するように更新する閉塞時ＲＡＩＤ構成情報更新手段と、
前記ディスクアレイ装置に対して閉塞状態にある前記ディスクアレイを使用可能な状態に回復させるリカバリ要求がユーザ操作に従って与えられた場合に、前記第２の記憶領域から、当該リカバリ要求で指定されたディスクアレイの閉塞直前の前記第２のＲＡＩＤ構成情報を読み出す手段と、
前記第２の記憶領域から読み出された前記第２のＲＡＩＤ構成情報に基づき前記ディスクアレイを閉塞直前の状態に強制的に戻すと共に当該第２のＲＡＩＤ構成情報を前記第１のＲＡＩＤ構成情報として前記第１の記憶領域に上書きする復旧手段とを備えることを特徴とするディスクアレイ装置。
【請求項７】　前記閉塞時ＲＡＩＤ構成情報更新手段は、前記閉塞手段により前記ディスクアレイが閉塞される際に、前記第２のＲＡＩＤ構成情報と共に当該ディスクアレイが閉塞した原因を示す情報を前記第２の記憶領域に保存するように構成され、
前記ディスクアレイ制御装置は、前記リカバリ要求が与えられた場合に前記読み出し手段によって前記第２の記憶領域から読み出された前記閉塞原因情報がディスクドライブの部分的な障害であるメディアエラーを示している場合、当該メディアエラーの発生していたブロックを代替用のセクタブロックに代替処理するセクタブロック代替手段を更に備えていることを特徴とする請求項６記載のディスクアレイ装置。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、ＲＡＩＤ（Redundant Arrays of Inexpensive Disks）構成のディスクアレイ装置に係り、特にディスクアレイを構成するメンバーのディスクドライブが複数故障した場合に当該ディスクアレイを使用可能な状態に回復するのに好適なディスクアレイのエラー回復方法、ディスクアレイ制御装置及びディスクアレイ装置に関する。
【０００２】
【従来の技術】一般にディスクアレイ装置は、複数のディスクドライブ、例えば磁気ディスクドライブ（以下、ＨＤＤと称する）から構成される少なくとも１つのディスクアレイと、このディスクアレイ内の各ＨＤＤ（メンバーＨＤＤ）に対するアクセスを制御するディスクアレイ制御装置とを備えている。
【０００３】ディスクアレイ装置は以下に述べる２つの

特徴を有する外部記憶装置として知られている。第１の特徴は、ホスト計算機上のファイルシステムから要求されたデータの読み出し／書き込みを、ディスクアレイ内の各ＨＤＤを並列に動かして分散して実行することでアクセスの高速化を図っている点である。第２の特徴は、データの冗長化によって信頼性の向上を図っている点である。

【０００４】ディスクアレイ制御装置は、ホスト計算機から転送される書き込みデータに対して、データ訂正情報としての冗長データを生成する。ディスクアレイ制御装置は、この冗長データをディスクアレイ内の複数のＨＤＤのうちのいずれか１つに書き込む。これにより、複数のＨＤＤのうちの１台が故障した場合、この冗長データと残りの正常なＨＤＤのデータを用いて故障したＨＤＤのデータを修復することが可能となる。

【０００５】データ冗長化の手法の１つとして、ＲＡＩＤ手法が知られている。ＲＡＩＤ手法は、ＲＡＩＤのデータと冗長データとの関連において、種々のＲＡＩＤレベルに分類される。ＲＡＩＤレベルの代表的なものにレベル３とレベル５がある。

【０００６】レベル３（ＲＡＩＤレベル３）では、ディスクアレイ制御装置は、ホスト計算機から転送される更新データ（書き込みデータ）を分割して、その分割された更新データ間の排他的論理和演算を行うことで冗長データとしてのパリティデータを生成する。ディスクアレイ制御装置は、このパリティデータで複数のＨＤＤのいずれかに書き込まれている元のパリティデータを更新する。一方、レベル５（ＲＡＩＤレベル５）では、ディスクアレイ制御装置は、ホスト計算機から転送される更新データ（新データ）と、当該更新データの格納先となるＨＤＤ内領域に格納されている更新前のデータ（旧データ）と、当該更新データの格納先に対応する別のＨＤＤの領域に格納されている更新前のパリティデータ（旧パリティデータ）との間の排他的論理和演算を行うことで、更新されたパリティデータ（新パリティデータ）を生成する。ディスクアレイ制御装置は、この新パリティデータで元のパリティデータを更新する。

【０００７】このようなＲＡＩＤ構成のディスクアレイ装置では、ディスクアレイ内のメンバーＨＤＤが故障した場合に、故障ＨＤＤのデータが次のように修復される。まずディスクアレイ制御装置は、故障したＨＤＤ以外の各ＨＤＤから、ディスクアレイのディスク領域を管理する単位であるストライプ毎にデータを読み出す。ディスクアレイ制御装置は、各ＨＤＤから読み出したデータの排他的論理和演算を行うことで、修復された（復元された）データを取得する。この排他的論理和演算を用いた手法、つまりＲＡＩＤのデータとパリティデータ（冗長データ）との整合性を利用したデータ修復を含む手法がＲＡＩＤ手法である。ディスクアレイ制御装置は、このＲＡＩＤ手法を用いて、修復されたデータをストライプ毎に取得することにより、故障したＨＤＤのすべての領域のデータを、故障したＨＤＤに代えて用いられるＨＤＤ内に修復することができる。この故障したＨＤＤに代えて用いられるＨＤＤは、故障したＨＤＤと交換して用いられるＨＤＤ、またはディスク制御装置に予め接続されていて、故障したＨＤＤの代替として割り付けられるスペアＨＤＤである。

【０００８】このように、ＲＡＩＤ構成のディスクアレイ装置では、ディスクアレイ内のメンバーＨＤＤが故障しても、故障したＨＤＤのデータを元通りに修復することができる。

【０００９】しかしながら、同一のディスクアレイ内で複数のメンバーＨＤＤが故障する、いわゆるＨＤＤの多重故障が発生した場合には、ＲＡＩＤのデータ冗長性を利用してユーザデータを修復することはできない。つまり、ＨＤＤの多重故障が発生した場合、ＲＡＩＤ手法によってユーザデータを修復することはできない。この場合、ディスクアレイ全体の故障となり、そのアレイは閉塞して、そのアレイ内のデータへはアクセスすることができなくなる。

【００１０】このようにディスクアレイが閉塞した場合、従来は故障したすべてのＨＤＤを交換し、改めてディスクアレイの状態を初期化した後に、別途データのバックアップが採取されたテープデバイスなどから、このアレイに対しデータを書き戻すのが一般的であった。

【００１１】

【発明が解決しようとする課題】上記したように、ＲＡＩＤ構成の従来のディスクアレイ装置では、ディスクアレイが閉塞した場合、故障したすべてのＨＤＤを交換し、改めてディスクアレイの状態を初期化した後に、データのバックアップが採取されたテープデバイスなどから、このアレイに対しデータを書き戻す必要があった。

【００１２】しかしながら、テープデバイスからの書き戻しには長時間を要するため、その間システムが停止状態となる問題があった。また、万一データのバックアップがなかった場合には、システムの復旧には膨大な時間と労力を必要とした。また、データのバックアップがあった場合でも、テープデバイスからの書き戻しには長時間を要するため、その間システムが停止状態となる問題もあった。

【００１３】本発明は上記事情を考慮してなされたものでその目的は、同一ディスクアレイ内でのディスクドライブの多重故障により当該ディスクアレイが閉塞したとしても、ＨＤＤ故障が一過性または部分的なものであったならば、外部からの要求に応じて当該ディスクアレイを簡単に使用可能な状態に復旧することができるディスクアレイのエラー回復方法、ディスクアレイ制御装置及びディスクアレイ装置を提供することにある。

【００１４】

【課題を解決するための手段】本発明に係るディスクア

レイのエラー回復方法は、複数のディスクドライブから構成されるＲＡＩＤ構成の少なくとも１つのディスクアレイを備えると共に、現在のディスクアレイの少なくとも構成及び稼働状態を示すＲＡＩＤ構成情報（第１のＲＡＩＤ構成情報）が保存される不揮発性のカレント用記憶領域（第１の記憶領域）と、ディスクアレイの閉塞時に、その閉塞直前の当該ディスクアレイのリカバリ用ＲＡＩＤ構成情報（第２のＲＡＩＤ構成情報）が保存されるリカバリ用記憶領域（第２の記憶領域）とが確保されたディスクアレイ装置におけるディスクアレイのエラー回復方法であって、ディスクアレイを構成する複数のディスクドライブのうちの少なくとも２つのディスクドライブが故障した場合に、当該ディスクアレイを閉塞するに際し、カレント用記憶領域に保存されている当該ディスクアレイの閉塞直前のＲＡＩＤ構成情報をリカバリ用ＲＡＩＤ構成情報としてリカバリ用記憶領域に保存すると共に、当該ディスクアレイの閉塞後に、当該ディスクアレイのカレント用記憶領域上のＲＡＩＤ構成情報を当該ディスクアレイの閉塞後の状態を反映するように更新し、閉塞状態にあるディスクアレイを使用可能な状態に回復させるリカバリ要求がユーザ操作により与えられた場合に、リカバリ用記憶領域から回復対象となるディスクアレイの閉塞直前のリカバリ用ＲＡＩＤ構成情報を読み出して、当該ＲＡＩＤ構成情報に基づき上記回復対象となるディスクアレイを閉塞直前の状態に強制的に戻すと共に当該リカバリ用ＲＡＩＤ構成情報をカレント用記憶領域に上書きすることを特徴とする。

【００１５】このように本発明においては、ディスクアレイ内の複数のディスクドライブの故障（つまりディスクドライブの多重故障）により当該アレイが閉塞する際、閉塞直前のＲＡＩＤ構成情報をリカバリ用ＲＡＩＤ構成情報としてリカバリ用記憶領域に保存し、当該アレイの閉塞後に、カレント用記憶領域上のＲＡＩＤ構成情報を当該アレイの閉塞後の状態を反映するように更新する。そして、ユーザの操作により閉塞状態にあるディスクアレイを対象とするリカバリ要求が与えられた場合、上記リカバリ用ＲＡＩＤ構成情報を読み出して当該ＲＡＩＤ構成情報に従い、指定されたディスクアレイを閉塞直前の構成や稼働状態に強制的に戻し、また当該ＲＡＩＤ構成情報をカレント用記憶領域に上書きする。

【００１６】これにより、ディスクドライブの多重故障にて対応するディスクアレイが閉塞されていても、緊急的にそのアレイの稼働を再開したい、または重要なデータだけでもそのアレイ内から読み出してバックアップを取りたいなどの理由で、ユーザの操作によりリカバリ要求が与えられた場合、ディスクドライブの故障が当該ドライブの電源をＯＦＦ／ＯＮしたり、モジュールを抜き差ししたりすることにより回復してしまう一過性のものであったならば、そのアレイを簡単に使用可能な状態に復旧することが可能となる。

【００１７】さて、ディスクドライブの故障には、上記一過性のものの他に、一部分の領域でのみ発生する障害（部分的な障害）、例えばメディアエラーがある。ディスクドライブの故障がメディアエラーに関する故障の場合、そのメディアエラーの発生していたセクタブロックを代替用のセクタブロックに代替処理してから、上記した指定ディスクアレイを閉塞直前の構成や稼働状態に強制的に戻すリカバリ処理を行うとよい。そのためには、ディスクアレイを閉塞するに際し、ＲＡＩＤ構成情報以外に当該ディスクアレイが閉塞した原因を示す情報もリカバリ用記憶領域に保存しておき、当該ディスクアレイのリカバリ処理を行う際にこの閉塞原因を示す情報も同時に読み出して、閉塞原因がメディアエラーにあるか否かを判定し、メディアエラーであるなら、上記のようにディアエラーの発生していたセクタブロックを代替用のセクタブロックに代替処理してからリカバリ処理を行うとよい。

【００１８】このようにすると、ディスクアレイ内の殆どのユーザデータは回復され、システムの稼働を継続することが可能となる。

【００１９】

【発明の実施の形態】以下、本発明の実施の形態につき図面を参照して説明する。図１は本発明の一実施形態に係るディスクアレイ装置の構成を示すブロック図である。

【００２０】図１において、ディスクアレイ装置１０は、少なくとも１つのディスクアレイ、例えば２つのディスクアレイ１１-1（＃１），１１-2（＃２）と、このディスクアレイ１１-1（＃１），１１-2（＃２）を制御するディスクアレイ制御装置１２とから構成される。

【００２１】各ディスクアレイ１１-1，１１-2は、いずれも複数のＨＤＤ（磁気ディスクドライブ）、例えば５台のＨＤＤ（メンバーＨＤＤ）１１１から構成される。

【００２２】ディスクアレイ制御装置１２は、ディスクアレイ装置１０を利用するホスト計算機２０とインタフェース２１を介して接続するための複数のインタフェースコントローラ１２１と、ディスクアレイ１１-1，１１-2を構成する各ＨＤＤ１１１とインタフェース２２を介して接続するための同数のインタフェースコントローラ１２２と、ホスト計算機２０とディスクアレイ装置１０との間で転送されるデータを一時的に格納するバッファメモリ１２３と、これらインタフェースコントローラ１２１及び１２２とバッファメモリ１２３との間のデータ転送に用いられるデータバス１２４とを備えている。インタフェース２１及び２２は、例えばＳＣＳＩ（Small Computer System Interface）、或いはファイバチャネル（FibreChannel）である。バッファメモリ１２３は、ディスクアレイ１１-1，１１-2の一部のデータの写しが保持されるディスクキャッシュを含んでいる。データバス１２４は、例えばＰＣＩ（Peripheral Component Int

erconnect Bus）である。

【００２３】ディスクアレイ制御装置１２はまた、当該制御装置１２の主制御部をなし、当該制御装置１２全体とディスクアレイ１１-1，１１-2とを制御するマイクロプロセッサ１２５と、マイクロプロセッサ１２５が実行する制御プログラム等が格納される不揮発性メモリ、例えばフラッシュメモリ１２６と、マイクロプロセッサ１２５のワークエリア等を提供する揮発性メモリ、例えばＲＡＭ１２７と、ディスクアレイ制御装置１２の状態表示及びディスクアレイ制御装置１２に対するユーザ操作による指示入力等に用いられる操作パネル１２８とを備えている。

【００２４】フラッシュメモリ１２６には、リカバリテーブル１２６ａとカレントテーブル１２６ｂとの２つのテーブルが置かれる。リカバリテーブル１２６ａはディスクアレイ１１-iが閉塞した際に、後で当該アレイ１１-iを使用可能な状態に戻す際の情報を保存するのに用いられる。この情報は、ディスクアレイ１１-iについての閉塞直前のＲＡＩＤ構成情報とアレイ１１-iが閉塞した原因を示す情報とから構成される。カレントテーブル１２６ｂは、現在のディスクアレイ１１-iの状態を示すＲＡＩＤ構成情報を保存するのに用いられる。

【００２５】ディスクアレイ１１-iのＲＡＩＤ構成情報は、

（１）ディスクアレイ１１-iで適用されるＲＡＩＤレベル、ストライプサイズ、論理容量（ユーザに提供されるディスクアレイ１１-i全体のディスク容量）を含むＲＡＩＤ基本情報

（２）ディスクアレイ１１-iを構成するＨＤＤ（メンバーＨＤＤ）１１１に関する情報（ＨＤＤ番号及び台数を含む）

（３）ディスクアレイ１１-iの稼働状態（アレイ全体として正常であるか閉塞しているか）と各メンバーＨＤＤ１１１の稼働状態（正常／故障）とを示す情報から構成される。

【００２６】ディスクアレイ制御装置１２は、複数のＨＤＤ（メンバーＨＤＤ）１１１によって構成されるディスクアレイ１１-i（ｉ＝１，２）のディスク領域を、図２に示すように複数のストライプ２０１に分割して管理する。

【００２７】ストライプ２０１は、ＲＡＩＤの手法により冗長性をもってストライピング配置された基本単位である。ストライピング配置とは、連続したデータを連続したディスクアレイ１１-iに順次マッピングすることをいう。冗長性とは、ストライプ２０１に含まれる各ＨＤＤ１１１の1つに、他の全ＨＤＤ１１１のデータの排他的論理和値が冗長データとして格納されていることをいう。この冗長性により、ディスクアレイ１１-1を構成する各ＨＤＤ１１１の1つが故障しても、ストライプ２０１に含まれる他の全ＨＤＤ１１１のデータの排他的論理和をとることにより、故障したＨＤＤ１１１の修復されたデータを生成することができる。ストライプ２０１のサイズは、１ＨＤＤ１１１当たり６４Ｋ（キロ）バイト～２５６Ｋバイト程度に設定されるのが一般的である。

【００２８】ストライプ２０１内のデータは、ＨＤＤ１１１の最小アクセス単位であるセクタブロック２０２を単位にアクセスされる。セクタブロック２０２のサイズは５１２バイトであるのが一般的である。

【００２９】各ＨＤＤ１１１には、セクタブロック２０２がエラー（メディアエラー）となった場合、つまり不良セクタとなった場合に、当該不良セクタの代替用として用いられる専用のセクタブロック、いわゆる代替ブロック（代替セクタブロック）２０３が複数確保されている。

【００３０】次に、図１の構成のディスクアレイ装置１０における動作について、（Ａ）ホスト計算機２０からのデータアクセス要求受信時の動作、（Ｂ）ＨＤＤ故障検出時の動作、（Ｃ）ユーザ操作によるディスクアレイのリカバリ（復旧）要求の受信時の動作を例に、順次説明する。

【００３１】（Ａ）データアクセス要求受信時の動作

まず、ホスト計算機２０からディスクアレイ装置１０に対してデータアクセス要求が発行された場合の当該ディスクアレイ装置１０の動作を、図３のフローチャートを参照して説明する。

【００３２】ホスト計算機２０からインタフェース２１を介してディスクアレイ装置１０に発行されたデータアクセス要求は、ディスクアレイ制御装置１２内のインタフェースコントローラ１２１で受け取られて、当該ディスクアレイ制御装置１２内のマイクロプロセッサ１２５に渡される。

【００３３】マイクロプロセッサ１２５は、ホスト計算機２０からのデータアクセス要求を受け取ると、アクセスの対象となるディスクアレイ（対応ディスクアレイ）１１-i（ｉは1または２）が稼働中であるか或いは閉塞中であるかを判定する（ステップＳ１）。この判定は、フラッシュメモリ１２６上に確保されたカレントテーブル１２６ｂに保存されている対応ディスクアレイ１１-iのＲＡＩＤ構成情報（に含まれている当該ディスクアレイ１１-iの稼働状態を示す情報）に基づいて行われる。

【００３４】もし、対応ディスクアレイ１１-iが稼働中の場合、マイクロプロセッサ１２５は当該ディスクアレイ１１-iへのデータアクセス処理を起動する（ステップＳ２）。これに対し、対応ディスクアレイ１１-iが閉塞中の場合には、マイクロプロセッサ１２５は当該ディスクアレイ１１-iに対するアクセスを行わずに、エラー終了する（ステップＳ３）。その理由は、次の通りである。

【００３５】まず、同一ディスクアレイ１１-i内で複数のＨＤＤ（メンバーＨＤＤ）１１１が故障するＨＤＤの

多重故障が発生した場合、ＲＡＩＤの冗長性を利用したデータ修復が不可能となるため、当該ディスクアレイ１１-iは閉塞する。もし、ＨＤＤの多重故障が発生したディスクアレイ１１-iを閉塞せずに、使用可能な状態に継続すると、故障したＨＤＤへのアクセスにより他の正常なＨＤＤへのアクセスに悪影響（ＨＤＤのインタフェースがロックするなど）を与える可能性がある。そのため、ＨＤＤの多重故障が発生したディスクアレイ１１-iは閉塞し、つまり他の正常なＨＤＤが存在しても、ディスクアレイ１１-i全体としては故障であるとして、以降アクセスしないようにしている。

【００３６】（Ｂ）ＨＤＤ故障検出時の動作

次に、ディスクアレイ１１-iを構成するＨＤＤ１１１の故障を検出した場合の動作を、図４のフローチャートを参照して説明する。

【００３７】マイクロプロセッサ１２５は、フラッシュメモリ１２６に格納されている制御プログラムに従い、各ディスクアレイ１１-i中の各ＨＤＤ１１１を定期的に監視する。ここでは、各ＨＤＤ１１１の記憶内容を読み出すことにより当該ＨＤＤ１１１の部分的な障害（メディアエラー）を検出するメディア検査処理が、閉塞中のディスクアレイ１１-iを構成するＨＤＤ１１１も対象として行われる。

【００３８】さて、上記メディア検査処理、或いは先のデータアクセス処理等で新たにＨＤＤ１１１の故障が検出された場合、マイクロプロセッサ１２５は、検出されたＨＤＤ１１１をメンバーＨＤＤとする対応するディスクアレイ１１-iが稼働中であるか或いは閉塞中であるかを判定する（ステップＳ１１）。

【００３９】もし、対応ディスクアレイ１１-iが稼働中である場合、マイクロプロセッサ１２５は新たにメンバーＨＤＤ１１１の故障が検出されても、当該ディスクアレイ１１-iはまだ稼働可能であるか否かを判定する（ステップＳ１２）。ここでは、対応ディスクアレイ１１-i内の他の全てのＨＤＤ１１１が正常である場合に、当該ディスクアレイ１１-iは稼働可能と判定される。これに対し、対応ディスクアレイ１１-i内の他の少なくとも１つのＨＤＤ１１１が既に故障である場合、つまり新たなＨＤＤ１１１の故障により、ＲＡＩＤの冗長性をもってしても故障したＨＤＤ１１１データが修復不可能な多重の故障状態となった場合、当該ディスクアレイ１１-iは稼働不可能であり、当該ディスクアレイ１１-iを閉塞する必要があると判定される。

【００４０】マイクロプロセッサ１２５は、対応ディスクアレイ１１-iが稼働不可能であると判定した場合、当該ディスクアレイ１１-iを閉塞するに際し、後で当該ディスクアレイ１１-iの状態を使用可能な状態に戻す際の情報として、現在の、つまり閉塞直前のＲＡＩＤ構成情報と閉塞原因とを、フラッシュメモリ１２６上のリカバリテーブル１２６ａに保存する（ステップＳ１３）。

【００４１】次にマイクロプロセッサ１２５は、対応ディスクアレイ１１-iを閉塞する処理を行う（ステップＳ１４）。このときマイクロプロセッサ１２５は、ディスクアレイ１１-i内の複数のＨＤＤ１１１の故障により、ＲＡＩＤのデータ冗長性をもってしても故障したＨＤＤ１１１内のデータを修復することが不可能となり、その結果当該ディスクアレイ１１-iを閉塞したことを、操作パネル１２８またはホスト計算機２０上の専用ソフトウエアを通して通知する。そしてマイクロプロセッサ１２５は、フラッシュメモリ１２６上のカレントテーブル１２６ｂに現在保存されている対応ディスクアレイ１１-iのＲＡＩＤ構成情報を、ステップＳ１４でのアレイ閉塞を反映した新たなＲＡＩＤ構成情報に更新し（ステップＳ１５）、しかる後にステップＳ１６に進む。

【００４２】一方、ステップＳ１１で対応ディスクアレイ１１-iが閉塞中であると判定された場合、或いはステップＳ１２でディスクアレイ１１-iが稼働可能であると判定された場合、マイクロプロセッサ１２５はそのままステップＳ１６に進む。

【００４３】マイクロプロセッサ１２５はステップＳ１６において、上記新たに故障が検出されたＨＤＤ１１１を閉塞する（ディスクアレイ１１-iから切り離す）処理を行う。そしてマイクロプロセッサ１２５は、カレントテーブル１２６ｂに現在保存されている対応ディスクアレイ１１-iのＲＡＩＤ構成情報を、ステップＳ１６での故障ＨＤＤ１１１の閉塞を反映した新たなＲＡＩＤ構成情報に更新する（ステップＳ１７）。

【００４４】（Ｃ）ディスクアレイのリカバリ要求受信時の動作

次に、閉塞中のディスクアレイ１１-iを対象とするリカバリ（回復）処理が要求された場合の動作について、図５のフローチャートを参照して説明する。

【００４５】ユーザにとって、ＨＤＤの多重故障にてディスクアレイが閉塞した場合においても、緊急的にそのアレイの稼働を再開したい、または重要なデータだけでもそのアレイ内から読み出してバックアップを取りたいことがある。このような場合、本実施形態ではディスクアレイ装置１０のディスクアレイ制御装置１２に設けられた操作パネル１２８をユーザが操作することで、所望のディスクアレイ１１-iのリカバリを要求することが可能なようになっている。また本実施形態では、ディスクアレイ装置１０を利用するホスト計算機２０からも、当該ホスト計算機２０にインストールされた専用のソフトウェアに従い、ユーザの操作に応じてホスト計算機２０からディスクアレイ装置１０のディスクアレイ制御装置１２に対してディスクアレイ１１-iのリカバリを要求することが可能なようにもなっている。

【００４６】さて、ユーザ操作により操作パネル１２８またはホスト計算機２０から発行されたディスクアレイリカバリ（復旧）要求はディスクアレイ制御装置１２の

マイクロプロセッサ１２５で受け付けられる。マイクロプロセッサ１２５は、このリカバリ要求を受け付けると、当該要求で指定されたディスクアレイ（対応ディスクアレイ）１１-iが閉塞中であるか否かを、カレントテーブル１２６ｂ上の該当するＲＡＩＤ構成情報に基づいて判定する（ステップＳ２１）。

【００４７】もし、対応ディスクアレイ１１-iが閉塞中でないならば、マイクロプロセッサ１２５はそのままリカバリ要求に対するリカバリ処理を終了する。

【００４８】これに対し、対応ディスクアレイ１１-iが閉塞中であるならば、マイクロプロセッサ１２５はリカバリテーブル１２６ａから、対応ディスクアレイ１１-iが閉塞する直前のＲＡＩＤ構成情報及び閉塞した原因を読み出す（ステップＳ２２）。そしてマイクロプロセッサ１２５は、閉塞の原因が、対応ディスクアレイ１１-iの閉塞直前に故障となったＨＤＤ１１１（つまり対応ディスクアレイ１１-iを構成するＨＤＤ１１１のうち、閉塞のトリガとなったＨＤＤ１１１）におけるセクタブロックのメディアエラー（部分的な障害）にあるか否かを判定する（ステップＳ２３）。

【００４９】もし、対応ディスクアレイ１１-iの閉塞の原因が、セクタブロックのメディアエラーにあった場合、マイクロプロセッサ１２５は、当該セクタブロックのデータの読み出しテストを行う（ステップＳ２４）。そしてマイクロプロセッサ１２５は、この読み出しテストにより、依然として読み出しが不可能であるか否か、つまりメディアエラーが再現するか否かを判定する（ステップＳ２５）。

【００５０】もし、ディスクアレイ１１-iが閉塞する原因となったセクタブロックのメディアエラーが再現した場合、マイクロプロセッサ１２５は当該メディアエラーが再現したセクタブロックのデータを破棄し、当該セクタブロックを代替ブロックへ代替する代替処理を行う（ステップＳ２６）。この代替処理により、メディアエラーが再現したセクタブロックへのアクセスは正しく行われる。但し、以前のデータは消失している。そこでマイクロプロセッサ１２５は、セクタブロックのメディアエラーにより部分的なデータ消失が発生したことをユーザに通知する（ステップＳ２７）。この通知には、操作パネル１２８を通して通知する方法、或いはホスト計算機２０上の上記専用ソフトウエアを通して通知する方法が適用可能である。

【００５１】マイクロプロセッサ１２５はステップＳ２７を実行すると、先にステップＳ２２でリカバリテーブル１２６ａから読み出した閉塞直前のＲＡＩＤ構成情報を、カレントテーブル１２６ｂに上書きし（ステップＳ２８）、以降このＲＡＩＤ構成情報に従って動作する。

【００５２】一方、上記ステップＳ２３で対応ディスクアレイ１１-iの閉塞原因がメディアエラー以外にあると判定された場合、或いは上記ステップＳ２５で当該メディアエラーが再現しないと判定された場合、マイクロプロセッサ１２５はそのまま上記ステップＳ２８に進んで、ステップＳ２２で読み出した閉塞直前のＲＡＩＤ構成情報を、カレントテーブル１２６ｂに上書きする。

【００５３】以上により、一度閉塞したディスクアレイ１１-iが、閉塞直前の使用できていた状態に復旧される。これによりディスクアレイ１１-iが再度使用可能な状態となるため、緊急的に運用稼働を継続したり、そのディスクアレイ１１-i内の重要なデータのバックアップを採取したりすることが可能となる。但し、このような状態に真に復旧できるのは、あくまで閉塞の原因となったＨＤＤ１１１の故障が一過性の場合や、メディアエラーのように部分的な障害の場合のみである。恒久的なＨＤＤ１１１故障の場合は、閉塞直前のＲＡＩＤ構成情報を用いて対応ディスクアレイ１１-iを閉塞直前の状態に戻しても、原因となったＨＤＤ１１１の故障が取り除かれないため、当該ＨＤＤ１１１の故障が再び検出される。この場合、図４のフローチャートに従う処理により、対応ディスクアレイ１１-iは直ちに再度閉塞される。

【００５４】以上に述べた実施形態では、リカバリテーブル１２６ａ及びカレントテーブル１２６ｂがディスクアレイ制御装置１２に実装されたフラッシュメモリ１２６（書き換え可能な不揮発性メモリ）上に確保されているものとして説明したが、これに限るものではない。例えば、ＨＤＤ１１１内の一部領域を、ユーザデータの格納用とは別に、リカバリテーブル１２６ａ及びカレントテーブル１２６ｂの領域を含む、ディスクアレイ装置１０自身の管理情報の記憶用領域として割り当てるようにしてもよい。

【００５５】なお、本発明は、上記実施形態に限定されるものではなく、実施段階ではその要旨を逸脱しない範囲で種々に変形することが可能である。更に、上記実施形態には種々の段階の発明が含まれており、開示される複数の構成要件における適宜な組み合わせにより種々の発明が抽出され得る。例えば、実施形態に示される全構成要件から幾つかの構成要件が削除されても、発明が解決しようとする課題の欄で述べた課題が解決でき、発明の効果の欄で述べられている効果が得られる場合には、この構成要件が削除された構成が発明として抽出され得る。

【００５６】

【発明の効果】以上詳述したように本発明によれば、ディスクドライブの多重故障により当該ディスクドライブをメンバーとするディスクアレイが閉塞してアクセス不能な状態となった場合においても、当該ディスクアレイを閉塞する際に、当該ディスクアレイについての閉塞直前のＲＡＩＤ構成情報をリカバリ用ＲＡＩＤ構成情報として不揮発性のリカバリ用記憶領域に保存しておくことにより、その後、閉塞したディスクアレイのリカバリ

（復旧）がユーザ操作に従って要求された場合に、リカバリ用記憶領域に保存されているリカバリ用ＲＡＩＤ構成情報に基づいて当該ディスクアレイを閉塞直前の状態に簡単に復旧することがでる。これにより、ディスクアレイの閉塞を招いたディスクドライブの故障が一過性のものまたはメディアエラーのような部分的なものであった場合には、上記復旧後のディスクアレイをアクセスして緊急的に運用を継続したり、当該アレイ内の重要なデータのバックアップを採取することができ、ユーザデータの保護を図ると共にシステムに与える被害を最小限に抑えることができる。

【図面の簡単な説明】

【図１】本発明の一実施形態に係るディスクアレイ装置の構成を示すブロック図。

【図２】図１中のディスクアレイ１１-i（ｉ＝１，２）のディスク領域を管理するのに用いられるストライプ、ＨＤＤ１１１の最小単位であるセクタブロック及び当該セクタブロックが不良セクタとなった場合に、当該不良セクタの代替用として用いられる代替ブロックの関係を説明するための図。

【図３】データアクセス要求受信時の処理手順を説明するためのフローチャート。

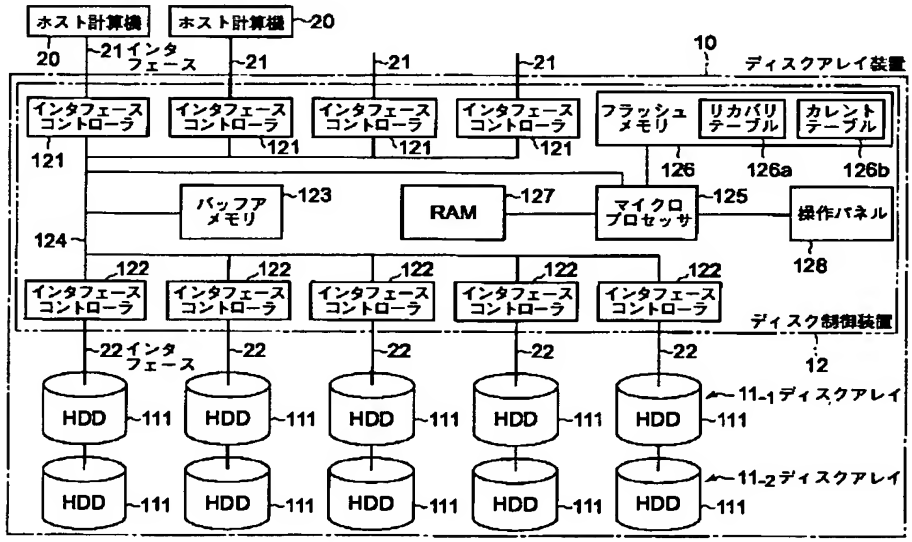
【図４】ＨＤＤ故障検出時の処理手順を説明するためのフローチャート。

【図５】リカバリ要求受信時の処理手順を説明するためのフローチャート。

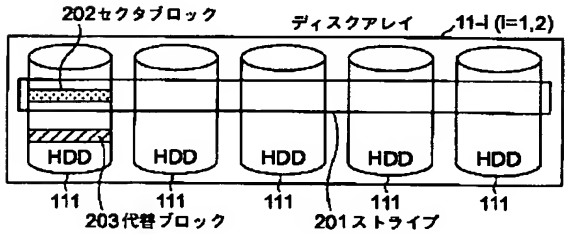
【符号の説明】

１０…ディスクアレイ装置
１１-1，１１-2，１１-i…ディスクアレイ
１２…ディスクアレイ制御装置
１１１…ＨＤＤ（ディスクドライブ）
１２５…マイクロプロセッサ
１２６…フラッシュメモリ
１２６ａ…リカバリテーブル（第２の記憶領域）
１２６ｂ…カレントテーブル（第１の記憶領域）
１２８…操作パネル

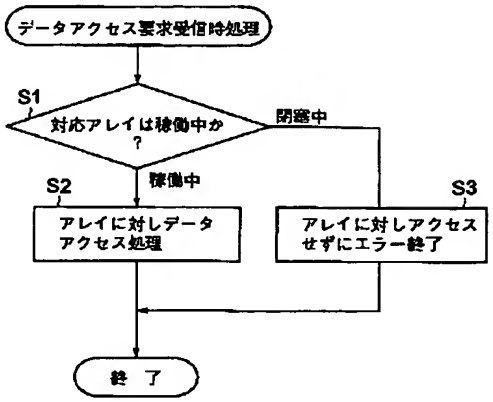
【図１】



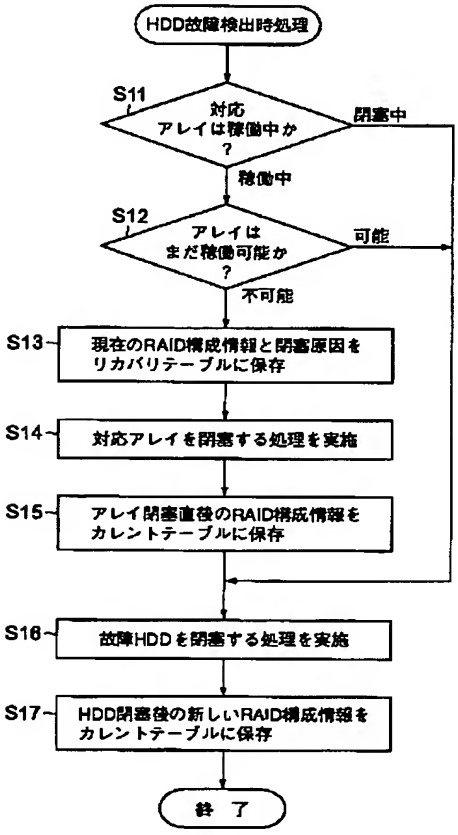
【図２】



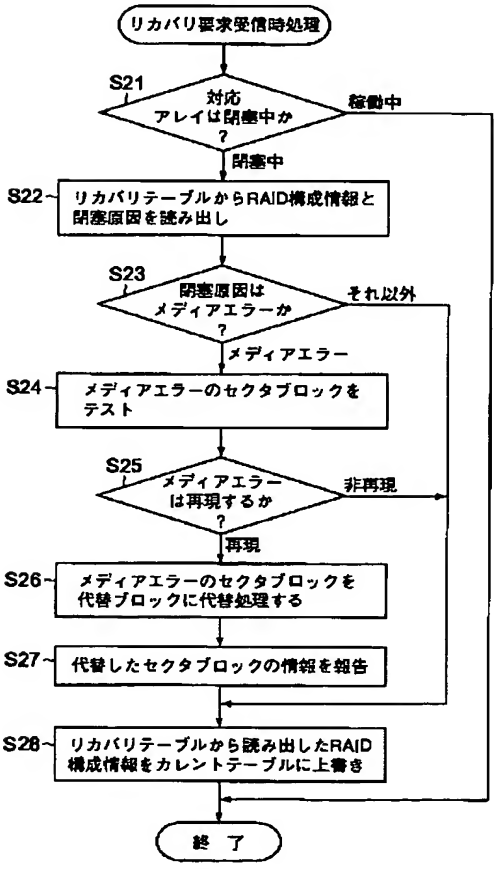
【図3】


データアクセス要求受信時処理
S1
対応アレイは稼働中か？
閉塞中
稼働中
S2
アレイに対しデータアクセス処理
S3
アレイに対しアクセスせずにエラー終了
終了

【図4】


HDD故障検出時処理
S11
対応アレイは稼働中か？
閉塞中
稼働中
S12
アレイはまだ稼働可能か？
可能
不可能
S13
現在のRAID構成情報と閉塞原因をリカバリテーブルに保存
S14
対応アレイを閉塞する処理を実施
S15
アレイ閉塞直後のRAID構成情報をカレントテーブルに保存
S16
故障HDDを閉塞する処理を実施
S17
HDD閉塞後の新しいRAID構成情報をカレントテーブルに保存
終了

【図5】


リカバリ要求受信時処理
S21
対応アレイは閉塞中か？
稼働中
閉塞中
S22
リカバリテーブルからRAID構成情報と閉塞原因を読み出し
S23
閉塞原因はメディアエラーか？
それ以外
メディアエラー
S24
メディアエラーのセクタブロックをテスト
S25
メディアエラーは再現するか？
非再現
再現
S26
メディアエラーのセクタブロックを代替ブロックに代替処理する
S27
代替したセクタブロックの情報を報告
S28
リカバリテーブルから読み出したRAID構成情報をカレントテーブルに上書き
終了